Quality Control for NGS Raw Data
Contents

- Data formats
  - Sequence capture
  - Fasta and fastq formats
  - Sequence quality encoding
- Quality Control
  - Evaluation of sequence quality
  - Quality control tools
  - Identification of artifacts & filtering
- Practical session
Sequence capture

RAW data
Propietary format

FastQ

Quality Control for NGS Raw Data

Francisco García
fgarcia@cipf.es
Genome sequencing

Reference genome

Quality Control for NGS Raw Data

Francisco García
fgarcia@cipf.es
Where are we?

Sequence preprocessing

Mapping

Variant calling

Variant prioritization

Functional annotation

GWAS Analysis

Gene-Set Analysis

Francisco García
fgarcia@cipf.es

Quality Control for NGS Raw Data
From sequencers to digital data

- **What structure does the data have?**
  - Text-based formats (easy to use!)
  - If not compressed, it can be huge

- **Different data formats:**
  - Different sequencers output different files (sff, fasta, csfasta, qual file, fastq...)
  - There are some data formats widely accepted (e.g. FastQ format)

Francisco García
fgarcia@cipf.es
Fasta format

- Two lines per sequence:
  1. Header lines starts with “>” followed by a sequence ID
  2. Sequence (string of nt or peptides)

```
>gi|5524211|gb|AAD44166.1|  cytochrome b [Elephas maximus maximus]
LCLYTHIGRIYYSYLQSEWTNGTHMLIIIIMATAFMYVLPWGQMSFWGATVITNLFAIPYIAYLTNLV
EWIWGGFSVSDKATLNRFFFAQHFILPFTMVALAGVHLTFLHDTGSNPLGLTSSDKIPFHYPYTIKDFLG
LLILILLLLALLSPDLGDPDNHPADPLNTPHIKPEWYFLFAYAILRSVPNKLGGVLALFSLIVIL
GLMPFLHTSKHRSMMLRPLSQLFWTLTMDDLTLTWIGSQPVEYPYTIIGQMASYLIFYSIILAPIAGX
IENY

>BBTBSCRYR
tgcaccaaacatgtctaaagctggaaccaaaattactttctttgaagcacaacacttttca
agggccgcactatgacaccgacattgcagttcgagttccacatgtacctgagcggctg
caactccatcagagtgaaggagccacctgggctgtgtatgtaaaggcccaatattttagctgg
gtacatgtacatcctcaccgggggcagatctacctgccactggatgggctca
```

- Typical file extensions (.fasta, .fa, .fna, .fnn, .faa, ...)

Francisco García
fgarcia@cipf.es

Quality Control for NGS Raw Data
We could say “it is a fasta with qualities”:

- 1. Header (like the fasta but starting with “@”)
- 2. Sequence (string of nt)
- 3. “+” and sequence ID (optional)
- 4. Encoded quality of the sequence

```plaintext
@SEQ_ID
GATTTGGGGTTCAAGCAGTATCGATCAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((((+++))))%+%++)(%+%).1***-++'))**55CCF>>>>>>CCCCCCC65
```
Quality codification

- **Phred quality score**
- **Error probability**
- **ASCII encoded**

**Phred +33**
- Sanger [0,40]
- Illumina 1.8 [0,41]
- Illumina 1.9 [0,41]

**Phred +64**
- Illumina 1.3 [0,40]
- Illumina 1.5 [3,40]

Francisco García
fgarcia@cipf.es

Quality Control for NGS Raw Data

Source: www.LookupTables.com
Quality codification

- **Phred scores**

\[ Q = -10 \log_{10} P \quad \leftrightarrow \quad P = 10^{-\frac{Q}{10}} \]

<table>
<thead>
<tr>
<th>Phred Quality Score</th>
<th>Probability of incorrect base call</th>
<th>Base call accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1000</td>
<td>99.9%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10000</td>
<td>99.99%</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100000</td>
<td>99.999%</td>
</tr>
</tbody>
</table>

Francisco García
fgarcia@cipf.es
If we evaluate our sequence in depth ... 
... we will know how reliable our results are

- **Problem:**
  - **Huge files** → Need of a tool to do it
Sequence quality evaluation

- Quality control tools:
  - Fastx-toolkit
    - [http://hannonlab.cshl.edu/fastx_toolkit/download.html](http://hannonlab.cshl.edu/fastx_toolkit/download.html)

- NGS QC Toolkit
  - [http://www.nipgr.res.in/ngsqctoolkit.html](http://www.nipgr.res.in/ngsqctoolkit.html)

Francisco García
fgarcia@cipf.es

Quality Control for NGS Raw Data
Sequence quality evaluation

- Other quality control tool: FastQC

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Francisco García
fgarcia@cipf.es

Quality Control for NGS Raw Data
Sequence quality per base position

- **Good data**
- **Consistent**
- **High quality along the read**

- The central red line is the median value
- The yellow box represents the inter-quartile range (25-75%)
- The upper and lower whiskers represent the 10% and 90% points
- The blue line represents the mean quality

Francisco García
fgarcia@cipf.es

Quality Control for NGS Raw Data
Sequence quality per base position

- Bad data
  - High variance
- Quality decrease with length

SOLiD characteristic pattern
Per sequence quality distribution

- Good data
- Most are high-quality sequences

Francisco García
fgarcia@cipf.es

Quality Control for NGS Raw Data
Per sequence quality distribution

- Bad data
- Non-uniform distribution

Quality score distribution over all sequences

Francisco García
fgarcia@cipf.es

Quality Control for NGS Raw Data
Per base sequence content

- Good data
- Smooth over length
- Organism dependent (GC)

Francisco García
fgarcia@cipf.es

Quality Control for NGS Raw Data
Per base sequence content

- Bad data
- Sequence position bias

Francisco García
fgarcia@cipf.es

Quality Control for NGS Raw Data
Per sequence GC content

- Good data
  - Fits with expected
  - Organism dependent

- Bad data
  - Does not fit with expected
  - Library contamination?
Per base N content

- **Good data**
- **Bad data**

Not good if there are N bias per base position

Francisco García
fgarcia@cipf.es

Quality Control for NGS Raw Data
Just descriptive:

Some sequencers output sequences of different length (e.g. 454)
**Sequence duplication levels**

- In **transcriptomics**, you expect higher number of duplicated sequences.
- In **genomics** you should be worried if this happens → PCR artifact?

---

Francisco García  
fgarcia@cipf.es

---

**Quality Control for NGS Raw Data**
Question:
- If we obtain the exact same sequences too many times
  → **Do we have a problem?**

Answer:
- Sometimes!

Examples:
- PCR primers, adapters ...

---

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGAGTTTTATCGCTTCCATGAC GCAGAAGTTAACACTTTC</td>
<td>2065</td>
<td>0.522403918155876</td>
</tr>
<tr>
<td>GATTGGCGTATCCAACCTGCA GAGTTTTATCGCTCCATG</td>
<td>2047</td>
<td>0.517850276254275</td>
</tr>
</tbody>
</table>
Typical artifacts

- Sequence adapters

---

Francisco García
fgarcia@cipf.es

---

Quality Control for NGS Raw Data
Typical artifacts

- Platform dependent

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

Francisco García
fgarcia@cipf.es
Filtering & trimming

- Removing bad quality data will improve our confidence on downstream analysis
Filtering & trimming

- Sequence filtering
  - Mean quality
  - Read length
  - Read length after trimming
  - Percentage of bases above Q
  - Adapter trimming
  - Adapter reads

 Francisco García
 fgarcia@cipf.es

Quality Control for NGS Raw Data
Filtering & trimming

- Sequence filtering tools
  - Fastx-toolkit
  - Galaxy (https://main.g2.bx.psu.edu/)
  - SeqTK (https://github.com/lh3/seqtk)
  - Cutadapt (http://code.google.com/p/cutadapt/)
  - And more....
Practical: FastQC & Fastx-toolkit

- Use FastQC to see your starting state.
- Use Fastx-toolkit to optimize different datasets and then visualize the result with FastQC to prove your success!

Hints: Try trimming, clipping and quality filtering.

Go to the tutorial and try the exercises...
Trimming the sequence with a minimum quality threshold of 20:

A

start ___________________________ end

17 20 19 18 15   Phred score

B

start ___________________________ end

17 20   Phred score